Corpus informatisés de français médiéval : contraintes sur leur constitution et spécificités de leurs apports[*]

Sophie PREVOST
Lattice (CNRS / UMR 8094)

Les linguistes qui ont pour objet d'étude un état de langue disparu ont toujours travaillé sur des textes. Il ne s'agit pas d'un choix, mais d'une obligation. En effet, contrairement aux linguistes qui s'intéressent à un état de langue contemporain, ils ne peuvent faire appel à leur intuition et/ou à d'éventuels informateurs. Il n'est donc pas question de recourir à des énoncés forgés, qui seraient non contrôlables : seuls les énoncés relevés dans les textes sont fiables. S'appuyer sur des données attestées ne signifie pas pour autant travailler « sur corpus » au sens où on l'entend désormais (et cela d'ailleurs quel que soit l'état de langue considéré). Pour cela il convient, d'une part de se fonder sur des données ordonnées, cohérentes, et aussi représentatives que possible au regard de l'objectif en fonction duquel elles ont été constituées, et d'autre part, bien souvent, de travailler sur des données numérisées, ce qui permet un traitement facilité à des degrés divers selon les outils dont on dispose.

Si l'élaboration de tout corpus est assujettie à certaines contraintes, en particulier celle de sa représentativité, le fait de travailler sur un état de langue disparu, en l'occurrence le français médiéval, renforce et complexifie cette exigence, qui varie en fonction d'au moins deux paramètres : le type d'étude (phénomène linguistique précis ou au contraire descriptif général) et la perspective temporelle (synchronique ou diachronique). On se penchera donc dans un premier temps sur

[*] Je tiens à remercier sincèrement les deux relecteurs anonymes de cet article pour leurs remarques très constructives, qui m'ont permis de rectifier ou de préciser certains points. Ceux qui peuvent sembler contestables au lecteur demeurent sous mon entière responsabilité.

les différentes contraintes qui pèsent sur la constitution d'un corpus, en essayant de proposer quelques éléments de réponse. On envisagera ensuite, à l'aide de quelques exemples, ce que l'usage raisonné des corpus, en « aval », apporte en plus, tant en ce qui concerne la quantification des faits que leur interprétation qualitative.

1. La constitution du corpus

1.1. Une langue sans locuteurs

Rappelons que travailler sur le français médiéval signifie travailler sur une langue sans locuteur, c'est-à-dire pour laquelle nous n'avons pas de compétence et ne pouvons faire appel à des informateurs qui l'auraient. Au mieux, une fréquentation assidue des textes, qui vient compléter et exemplifier une grammaire acquise par ailleurs (mais dont l'élaboration s'est néanmoins fondée sur l'exploration et l'analyse des textes) peut permettre de développer une certaine « intuition de reconnaissance »[1].

Seul accès à la langue, les textes sont donc essentiels[2]. Sachant qu'il n'est pas possible de tous les prendre en considération, la sélection que l'on opère pour construire un corpus (qu'il soit destiné à une étude spécifique ou générale) est décisive. Celui-ci constituera en effet la seule image de la langue, puisqu'en l'absence de véritable « intuition de locuteur » il n'y aura pas de garde-fou fiable aux analyses tirées des textes[3]. L'attention à accorder à la constitution du corpus est

[1] Marchello-Nizia (1995 : 22) suggère que le linguiste médiéviste peut parfois développer, à défaut d'une compétence de « production », une compétence de « reconnaissance ».

[2] Le linguiste médiéviste semble prisonnier d'une situation circulaire, étudiant les textes avec une connaissance de la langue fondée sur ces mêmes textes. Une démarche critique, largement fondée sur la prise en compte régulière de nouveaux textes, et la mise au jour concomitante de nouveaux faits, ou la rectification de précédentes analyses, permet cependant d'échapper à la circularité.

[3] Dans le cadre d'un travail sur un état de langue contemporain, l'intuition permet certes d'évaluer avec davantage de certitude le caractère grammatical/acceptable d'une construction, mais elle ne permet pas pour autant de dater précisément l'émergence (ou la disparition) d'une

d'autant plus grande si celui-ci doit servir à une étude dont les résultats sont destinés à être généralisés.

1.2. La représentativité du corpus

Quel que soit le type d'étude envisagée, le corpus utilisé doit être représentatif, et cela vaut d'ailleurs pour tout état de langue. Il est désormais acquis que la représentativité doit se décliner selon deux dimensions, quantitative et qualitative, l'antagonisme initial des deux ayant fait place à l'évidence de leur complémentarité[4] : l'accumulation des données, permise par leur numérisation croissante, perd en effet de son utilité si elle n'est pas complétée par leur diversification raisonnée.

1.2.1. Considérations quantitatives

Du point de vue quantitatif, il convient de déterminer, pour chaque catégorie textuelle retenue, le nombre de textes ou d'échantillons, ainsi que la taille en mots de ces derniers[5]. Les propositions qui ont été faites, en particulier par Biber[6], restent indicatives, susceptibles de varier selon le type d'étude. Il est en effet difficile d'envisager une norme absolue en matière de « taille », du fait qu'il n'existe pas de critères véritablement objectifs pour la fonder. L'essentiel demeure donc de veiller à la cohérence et à l'homogénéisation quantitatives internes du corpus.

---

construction (même récente) ni de fournir des fréquences fiables, et encore moins détaillées.

[4] Pour un historique de cette question, voir Péry-Woodley (1995).

[5] La question du nombre de mots ne se pose pas pour les textes intégraux, mais il convient en revanche de pondérer les résultats en fonction de la taille variable des textes. Notons à cet égard que la notion de texte intégral mériterait d'être discutée, tant les « unités textuelles » varient selon qu'il s'agit d'un proverbe, d'un lai, d'une chronique, ou bien d'un roman en plusieurs volumes.

[6] S'appuyant sur les caractéristiques du corpus LOB, Biber (1990) propose de retenir 10 textes par catégorie, et, se fondant sur la représentativité de nombreux traits grammaticaux, il suggère une moyenne de 1 000 mots par échantillon ; il admet cependant qu'un échantillon plus important est nécessaire pour certains traits plus rares.

Le choix de travailler sur des échantillons, pour les textes longs, répond généralement à des raisons pratiques : cela permet de traiter davantage de textes. Cette démarche est cependant risquée si l'étude implique la recherche de constructions rares, ou de formes dont l'existence demeure incertaine. Le linguiste médiéviste ne peut en effet s'appuyer sur son intuition de locuteur pour décider si la forme existe ou non : seuls les textes peuvent fournir une réponse, et le recours à de simples échantillons risque de faire passer à côté d'une occurrence dont on ne soupçonne pas l'existence. Le recours à l'échantillonnage, s'il a néanmoins lieu, doit s'accompagner de précautions, dans la mesure où les textes peuvent être hétérogènes de différents points de vue. Ils peuvent l'être sur le plan narratif[7], ou bien chronologique, lorsque la rédaction de l'œuvre a été longue et que la langue de l'auteur a pu se modifier au fil des années. C'est le cas des *Mémoires* de Commynes, œuvre dans laquelle certains aspects de la syntaxe du sujet ont évolué entre le livre 1 et le livre 7, respectivement rédigés en 1489 et en 1498. Il faut aussi envisager les cas où l'écriture s'est faite à plusieurs mains, avec un éventuel décalage dans le temps, comme pour le *Roman de la rose*, dont la première partie a été rédigée par G. de Lorris en 1225-1230, et la seconde par J. de Meun en 1269-1278[8].

1.2.2. Considérations qualitatives

En ce qui concerne la dimension qualitative, il convient en préalable de déterminer les genres existants, et parmi eux ceux que l'on suppose pertinents pour l'objet d'étude.

Si la classification des textes est réputée complexe pour le français moderne, comme en témoigne l'absence d'une typologie consensuelle, elle l'est bien davantage encore pour le français médiéval, et ce pour différentes raisons. La première

[7] Voir les remarques de Guillot (2003) à propos *Des cas des nobles hommes* (Boccace, traduit par Laurent de Premierfait), qui alterne plusieurs types narratifs.

[8] Cela rejoint la remarque faite en note 5.

tient à ce que les « genres[9] » de l'époque diffèrent en partie de ceux d'aujourd'hui. Par exemple les « romans » des 12ème et 13ème siècles n'ont guère en commun avec le roman moderne (si tant est qu'on puisse le définir de manière univoque). Certains écrits se laissent en outre difficilement classer, comme les textes « historiques » (chroniques, mémoires…), qui comprennent bien souvent une part de fiction (caractéristique des genres littéraires) et dénotent une implication assez forte de l'auteur. La classification des textes se révèle difficile aussi parce que nous manquons parfois de « matière » -de textes- pour caractériser un genre. Peut-être même ignorons-nous l'existence de certains genres, disparus sans avoir laissé de textes témoins[10]. A ces différentes difficultés s'ajoute le fait que des genres apparaissent, d'autres disparaissent, et surtout certains évoluent, ce qui est complexe à gérer dans la perspective d'une approche diachronique : faut-il considérer qu'il s'agit toujours du même genre, ou au contraire d'un autre ? Il n'est pas aisé de déterminer des critères objectifs pour mesurer le changement, et de décider du « seuil » éventuel au-delà duquel on doit considérer que l'on a affaire à un genre différent.

1.2.3. Disponibilité des textes

Ainsi, alors que c'est pour les états de langue disparus que nous aurions besoin de la meilleure représentativité possible sur les plans quantitatif et qualitatif, c'est précisément pour eux que cela se révèle le plus difficile. En effet, en admettant que l'on ait réussi à dresser la liste des genres, et à

---

[9] Nous adoptons ici une terminologie très simplifiée, en opposant genres littéraires (roman, nouvelle, chanson de geste, poésie, théâtre…) et genres non-littéraires (texte juridique, charte, coutumier…), qui se subdivisent en sous-genres. Une classification beaucoup plus fine et argumentée des textes a été élaborée à l'ENS-LSH de Lyon (C. Marchello-Nizia, C. Guillot et A. Lavrentiev). Elle distingue des « domaines » (religieux, littéraire, juridique, historique et didactique, et parmi eux des « genres » (hagiographie, épique, roman, lyrique, traité, serment, chronique, mémoires, encyclopédie…).

[10] Nous ne disposons de toute façon d'aucune trace d'oral véritable, au mieux pouvons-nous prendre en considération les « mises en écrit » de l'oral.

imaginer le corpus « idéal » pour l'étude à mener, se pose le problème de l'accès aux textes. Certains ont disparu, d'autres ne sont pas facilement utilisables car ils ne sont pas édités, ou s'ils le sont, pour un nombre encore élevé, ils ne sont pas numérisés[11]. D'une manière générale, il n'est pas toujours facile de réunir pour telle époque tel nombre de textes pour telle catégorie, *a fortiori* numérisés.

1.2.4. Risques liés à une mauvaise composition du corpus

L'équilibre du corpus et le bon ciblage des genres est pourtant essentiel, dans la mesure où toutes les structures ne sont pas également présentes dans les différents genres : en exclure certains, en sous-représenter d'autres fait courir deux risques majeurs, directement liés. Le premier consiste à conclure hâtivement et de manière erronée à l'absence d'une construction. Ainsi, dans le cadre d'une étude consacrée à l'évolution du marqueur de topicalisation *quant à* (Prévost 2003), la grande rareté des occurrences dans les textes d'ancien français de la BFM[12] nous avait d'abord fait penser qu'il s'agissait d'une forme peu utilisée à cette époque. Toutefois, l'examen des occurrences relevées pour le moyen français dans

[11] Même si les études diachroniques suscitent un intérêt croissant chez les collègues modernistes, le français médiéval reste encore un objet quelque peu marginal, ce qui explique le retard en matière de numérisation des textes, moins susceptibles de trouver des débouchés commerciaux immédiats. A cela s'ajoute la difficulté accrue de numériser ce genre de textes, en raison de l'absence de locuteur natif et de la variation graphique. Il faut aussi évoquer les fortes résistances de certains éditeurs commerciaux, qui bloquent la libre utilisation des textes, précisément lorsqu'il s'agit de les numériser. Malgré cela plusieurs bases d'une taille conséquente -plusieurs millions de mots- se sont peu à peu constituées, en particulier celles de la BFM et du DMF (voir notes 12 et 13).

[12] BFM : Base de français médiéval, UMR5191 ICAR / ENS-LSH Lyon (environ 3 millions de mots). Au moment où nous rédigeons cet article, le site de la base est provisoirement indisponible pour des raisons juridiques et nous ne sommes donc pas en mesure d'en communiquer l'adresse.

la base du DMF[13] a conduit à réviser cette interprétation. C'est en effet dans des textes non littéraires (didactiques, argumentatifs...) que nous avons rencontré la majorité des occurrences, et les textes d'ancien français consultés étaient au contraire tous littéraires et narratifs : il était donc peu probable de faire une collecte bien fructueuse (d'autant que l'expression, tous genres confondus, reste plus rare en ancien qu'en moyen français). Le second risque consiste à vouloir généraliser les résultats obtenus, alors que le corpus d'étude, trop peu représentatif, n'autorise pas à tirer de conclusions au-delà des textes qui le constituent.

Les risques liés à un corpus trop peu représentatif existent pour tout état de langue, mais lorsque l'on possède la compétence d'une langue, ou que l'on peut faire appel à des informateurs, il est plus facile de corriger des conclusions erronées. Ainsi, on peut ne pas trouver telle construction dans un corpus, mais l'avoir déjà entendue ou lue ailleurs. L'intuition du locuteur n'est cependant pas entièrement fiable, en particulier dès qu'il s'agit d'évaluer la fréquence d'une construction (cf. note 3), et cela d'autant plus qu'on le sollicite sur des genres et/ou des registres différents. Il n'est ainsi pas rare qu'une estimation soit largement contredite par la réalité des faits. Le danger est moindre de ce point de vue quand on travaille sur une langue ancienne, car on se fie rarement à son intuition ! Reste toutefois la tentation de généraliser trop largement les résultats obtenus à partir de l'étude d'un corpus. Le risque décroît d'autant que le corpus est représentatif sur les plans quantitatif et qualitatif, mais il demeure difficile d'indiquer une norme absolue, du fait que la constitution du corpus est largement conditionnée par la nature de l'étude à mener.

De ce point de vue, on retiendra deux facteurs principaux de variation : le caractère plus ou moins spécifique de l'objet d'étude, la dimension diachronique ou synchronique de l'étude.

---

[13] Base du DMF : Base du Dictionnaire de Moyen Français, UMR7118 ATILF / Nancy-2, < http://atilf.atilf.fr/dmf.htm (environ 7 millions de mots).

1.3. L'objet de l'étude

L'objet d'étude conditionne la construction du corpus à deux niveaux. En amont, il donne des pistes quant aux textes à retenir. Pour reprendre l'exemple des marqueurs de topicalisation, une connaissance minimale de la question conduit à exclure les textes d'ancien français, et les textes littéraires pour les débuts du moyen français, dans la mesure où ces expressions semblent s'être d'abord développées dans les textes non littéraires. On sait de même que le déterminant *ledit* (et ses différentes variantes) n'apparaît qu'en moyen français, et ne se rencontre d'abord que dans des textes administratifs et juridiques. Ces restrictions quant aux genres n'ont en revanche pas leur raison d'être pour l'étude de bon nombre de phénomènes linguistiques. Parfois c'est la seule période qui s'avère pertinente, tel fait n'apparaissant pas avant telle époque, ou ayant au contraire disparu à partir de telle autre. La prudence reste de mise pour les périodes charnières, qui permettent précisément de repérer l'apparition ou la disparition d'un phénomène linguistique.

Notons que l'on peut aussi vouloir retenir dans le corpus des textes dans lesquels on sait la représentation du phénomène étudié absente ou très rare : d'une part le constat de l'absence d'un phénomène -qui est aussi un fait linguistique- participe de l'étude de sa distribution et de sa répartition ; cela peut d'autre part permettre d'identifier les faits linguistiques qui remplacent éventuellement le phénomène quand il est absent.

Si tant est que l'on ait pu déterminer quel serait le corpus idéal pour l'étude à mener, on se heurte en aval à des considérations « pratiques ». La non-disponibilité de certains textes, évoquée précédemment, reste un problème de taille, mais il faut aussi évoquer les contraintes liées à la « gestion » des données. Il s'agit d'une part du repérage des occurrences du fait linguistique traité, et d'autre part de leur traitement.

1.3.1. Repérage et extraction des données

Cet aspect de la question recouvre différents cas de figure, variables selon la complexité de la construction et selon les outils plus ou moins sophistiqués dont on dispose pour en

extraire les occurrences. Parmi les nombreux cas possibles, on en distinguera quelques-uns, assez typiques. Le premier correspond au repérage d'une forme simple prédéfinie, pour laquelle la seule difficulté éventuelle réside dans la détermination préalable des différentes graphies possibles. La requête peut être aisément satisfaite, sans recours à des outils complexes. Ceux-ci peuvent en revanche être nécessaires s'il s'agit d'une forme prédéfinie, mais discontinue, comme par exemple les occurrences de « ne … pas/point/mie » : il faut un outil capable de gérer la discontinuité. Dans les deux cas précédents, on travaille sur des formes, et un corpus non enrichi est suffisant. C'est différent lorsque les requêtes portent sur des catégories morpho-syntaxiques, voire sur des structures syntaxiques (les séquences « sujet-verbe » par exemple), puisque cela suppose de disposer d'un corpus étiqueté du point de vue morpho-syntaxique ou syntaxique. Or le français médiéval, comme d'autres états de langues anciens, accuse de ce point de vue un retard assez net comparé au français moderne[14]. En effet, alors que le recours aux corpus enrichis est désormais un acquis pour ce dernier, ce n'est pas encore le cas pour le français médiéval, même si des progrès considérables ont été accomplis récemment, ou sont en cours, souvent dans le cadre de vastes projets[15]. Le retard, en voie de se combler, s'explique principalement par l'absence d'étiqueteur « prêt à l'emploi » (incluant grammaire et dictionnaire) conçu pour les différents états de langue de cette époque[16]. L'évolution de la

[14] Pour des raisons en partie analogues à celles évoquées à propos de la simple numérisation des textes (voir note 11).

[15] Depuis plusieurs années 5 textes (plus de 300000 mots) de la BFM sont étiquetés à l'aide d'un jeu de 60 étiquettes morpho-syntaxiques (cf. 2.3. infra). Parmi les vastes projets actuels on citera le projet « Modéliser le changement : les voies du français », dirigé par F. Martineau à Ottawa, et le projet ANR « Corpus représentatif des premiers textes français », dirigé par Céline Guillot (ENS-LSH de Lyon)

[16] Les procédures par apprentissage, assez performantes, supposent d'être déjà en possession d'un ou plusieur(s) texte(s) étiqueté(s), et par ailleurs de procéder à des vérifications systématiques, la variation pouvant générer davantage d'erreurs que pour le français moderne. A propos de cette démarche, voir Prévost & Heiden (2002) et Stein (2003).

langue ne simplifie évidemment pas la tâche, d'autant que, même dans ses phases de relative stabilité, la langue connaît une variation morphologique et syntaxique (relative souplesse de l'ordre des des mots) qui complique fort l'élaboration de règles. Au demeurant, la seule conception d'un jeu d'étiquettes consensuel s'avère une affaire complexe, ne serait-ce que parce que certaines catégories apparaissent tandis que d'autres disparaissent[17]. Les précieux apports des textes enrichis, que nous évoquerons plus bas, doivent inciter à œuvrer pour leur multiplication.

Si la facilité du repérage et de l'extraction des données est un aspect essentiel de l'exploitation des corpus, la question du traitement des données est importante aussi.

1.3.2. Traitement des données

Outre la taille du corpus de départ, évoquée précédemment, il faut aussi se pencher sur celle du « sous-corpus » que constituent les faits pertinents, puisque, en aval de leur extraction, se pose la question de leur traitement. La nature de l'étude est à cet égard largement décisive. Il y a ainsi une différence nette entre le simple classement des emplois d'une forme et l'analyse d'une structure complexe, par exemple sur le plan sémantico-pragmatique : le premier est plus rapide que la seconde, et à temps égal, on traitera un plus grand nombre de résultats dans le premier cas que dans le second. Entre les deux, de nombreuses configurations sont possibles. Ainsi, pour la seule étude de formes simples, nous avons rencontré deux cas de figure très différents. Nous avons en effet étudié, de l'ancien français au 16ème siècle et sur de gros corpus[18], l'adverbe « aussi »[19] en position initiale et le trinôme « les aucuns » / « d'aucun(s) » / « aucun(s) »)[20]. Pour le premier, une fois

[17] Pour un développement de ces différents points en corrélation avec la démarche menée au sein de la BFM, voir Prévost & Heiden (2002).

[18] La BFM pour l'ancien français, la base du DMF pour le moyen français, et celle de Frantext pour le 16ème siècle (5,6 millions de mots) .

[19] Prévost (1999).

[20] Ici abrégé par commodité en /aucun/. Cf. Prévost & Schnedecker (2004).

sélectionnées les seules formes en position initiale, le total s'élevait 696 occurrences. Il était en revanche de 10 444 occurrences pour /aucun/ ! On a là l'illustration typique de la différence qu'il peut y avoir, à partir d'un même corpus, entre les sous-corpus de faits pertinents. Les données n'ont évidemment pas pu être traitées de la même manière. Dans les deux cas, on a dégagé les caractéristiques morpho-syntaxiques, mais la dimension sémantico-pragmatique de l'étude a en revanche été envisagée selon des modalités différentes : de manière détaillée et exhaustive pour « aussi », de manière sélective pour /aucun/, pour lequel il n'était pas possible d'envisager une étude sémantico-référentielle détaillée sur l'ensemble du corpus de travail. On a donc réduit, soit le choix des formes mêmes (« les aucuns » et « d'aucuns » bénéficiant d'une description plus minutieuse que « aucun »), soit celui de leurs caractéristiques morphologiques (formes plurielles et pronominales par exemple), soit encore on a opéré une sélection générique parmi les textes.

Nous avons par ailleurs mené une étude sémantico-pragmatique portant sur les énoncés à sujet pronominal de 3ème personne (préverbal ou postverbal)[21]. L'étude a dans un premier temps été réduite à quelques textes d'ancien et moyen français (4 au total), les seuls 1247 énoncés à analyser représentant une lourde tâche. Le fait que la sélection se soit opérée en amont, au niveau du nombre de texte retenus, et non pas en aval au niveau des occurrences elles-mêmes, comme pour /aucun/, tient dans ce cas à la relative complexité du relevé des occurrences pertinentes[22]. Le niveau où s'opère la sélection, quand elle a lieu, n'est pas prédéfini : il varie en fonction d'un certain nombre de facteurs, parmi lesquels figurent aussi des considérations très « pratiques », qui ne sont toutefois pas illégitimes si elles sont dominées et guidées par des considérations scientifiques.

On ne peut décider d'une manière générale du seuil au-delà duquel la gestion exhaustive du corpus et de ses données n'est plus possible : un tel seuil n'existe pas de manière

[21] Prévost (2008a).

[22] Bien moins cependant que dans le cas de sujets nominaux…

absolue, tant il dépend de facteurs divers (nature de l'analyse, degré de finesse de celle-ci…). Il n'en est pas moins utile de garder à l'esprit les contraintes qu'impose le « quantitatif »[23].

1.4. La perspective temporelle : synchronique ou diachronique[24]

Lorsque l'on travaille sur le français médiéval, la question « temporelle » est complexe. Elle recouvre à la fois la représentation que l'on a de la langue, et la manière dont on construit l'objet textuel qui va nous servir à étudier la langue. La première s'appuyant sur les textes, on n'échappe évidemment pas complètement à une certaine circularité…

L'appellatif « français médiéval » ne recouvre à vrai dire aucune réalité langagière homogène, dans la mesure où le français médiéval s'étend sur 7 siècles (9ème -15ème siècles) ! On a coutume, au sein de cette vaste période, d'opérer une distinction entre ancien (9ème-13ème) et moyen français (14ème-15ème)[25]. Il ne s'agit pas pour autant de périodes parfaitement stables : en leur sein certains faits langagiers ont changé, tandis que d'autres se sont maintenus, puisque, comme le rappelle C. Marchello-Nizia (1995 : 7) : « […] tout état de langue est à la fois, nécessairement, 'transition' et 'stabilité', dans la mesure où toute langue naturelle change, continûment, tout en maintenant un équilibre indispensable à l'intercompréhension. » Envisager une stabilité linguistique à une période donnée ne vaut donc que pour tel ou tel fait langagier, mais non pour leur ensemble.

---

[23] Notons cependant que, face aux données volumineuses, des procédures textométriques, en particulier l'établissement de concordances bien ciblées, peuvent déjà faire émerger des résultats significatifs, avant une analyse manuelle plus fine.

[24] Nous reprenons ici assez directement ce que nous avons dit sur cette question dans Prévost (2005 : 152-153)

[25] Même si l'exacte délimitation entre les deux périodes n'est pas totalement consensuelle. Rappelons d'ailleurs que cette distinction s'appuie sur des critères non seulement linguistiques mais aussi politico-socio-culturels. Pour une présentation détaillée de cette question, voir Marchello-Nizia (1997a : 3-9).

C’est évidemment le phénomène étudié et la décision de l’appréhender dans une phase de stabilité ou d’évolution qui conditionne le choix de la période considérée, même si on peut à l’inverse décider de décrire, pour telle période, les faits linguistiques qui la caractérisent.

Si la période pertinente pour l’étude n’est pas simple à déterminer, le choix des textes qui vont la représenter ne l’est pas non plus, et cela indépendamment même de la diversité qualitative. En effet, qu’elle soit synchronique ou diachronique, la démarche se heurte à la difficile délimitation « chronologique ». Pour une approche synchronique, se pose ainsi la question de la taille de la fenêtre temporelle adéquate pour rendre compte d’une large période (ancien ou moyen français dans leur totalité par exemple). Faut-il envisager les données sur l’ensemble de l’intervalle, ou peut-on se contenter de coupes plus étroites mais supposées représentatives de la période ? Quelle est dans ce cas la taille de la bonne fenêtre ? Dix ans ? (c’est le choix fait par Martin et Wilmet pour la *Syntaxe du moyen français* (1980), pour laquelle ils ont retenu la décennie 1455-1465). Plus ? Moins ? La réponse dépend en partie du niveau d’analyse visé, les faits ne bougeant pas au même rythme dans les différents domaines (sémantique lexicale, morphologie ou syntaxe).

Même si une approche diachronique ne se réduit nullement à décrire des synchronies successives, c’est néanmoins une étape nécessaire : il faut donc déterminer à la fois la taille de celles-ci et leur fréquence : tous les dix ans ? tous les vingt ans ? A moins qu’il faille au contraire envisager une couverture exhaustive (par exemple : 1220-1230, 1230-1240, 1240-1250…)… En laissant une période dans l’ombre, on risque d’omettre des faits notables, en particulier lorsque l’enjeu est de dater l’apparition ou la disparition d’une forme, ou bien d’attester l’existence d’une forme rare.

Au-delà de la question de la taille des fenêtres temporelles et de leur fréquence, c’est la nature même de l’évolution qui est en cause, avec à cet égard deux positions majeures : l’une d’elle consiste à considérer que les

changements surviennent brusquement[26], tandis que l'autre les envisage comme inscrits dans un continuum, avec par exemple coexistence temporaire d'une forme nouvelle et d'une forme ancienne, la seconde, initialement dominante, devenant minoritaire avant, souvent, de disparaître ou de se refaire une jeunesse dans d'autres emplois. Les textes nous montrent que, au moins en morphologie et en syntaxe, c'est plutôt selon un mode continu qu'évoluent les faits langagiers. Cette réalité incite à réduire autant que possible les « ellipses » temporelles, qui pourraient occulter une phase importante du développement du phénomène étudié.

1.5. Cas particulier : l'élaboration d'une « grammaire »

Jusqu'ici nous avons plutôt considéré l'étude de faits langagiers spécifiques. Il convient maintenant d'envisager la constitution d'un corpus dans la perspective de l'étude de l'ensemble des faits langagiers, que ce soit d'un point de vue synchronique ou diachronique. L'élaboration d'une grammaire constitue un tel projet. Les grammaires de langue ancienne, quelles que soit la période et la perspective temporelle envisagées, se sont nécessairement toujours appuyées sur des textes, mais dans des proportions variables[27], limitées pour beaucoup par l'exploitation plus lente des textes qu'implique leur support papier. Aujourd'hui, les progrès considérables en matière de numérisation et le fait qu'il n'est plus possible d'envisager un tel projet sans l'asseoir sur un corpus conséquent conduisent à s'interroger sur la constitution « concrète » d'un tel corpus.

---

[26] C'est la conception du changement qui avait été développée dans un premier temps dans le cadre de la grammaire générative : selon Lightfoot (1979), un changement linguistique ne peut être que 'catastrophique', dans la mesure où il suppose une *réanalyse*, et que seuls les enfants sont capables d'accomplir une telle opération, en analysant différemment les énoncés entendus. Les changements linguistiques sont donc liés aux changements générationnels. Lightfoot a par la suite adopté une position moins radicale.

[27] Voir à ce propos Prévost (2005, note 5, p. 149)

L'élaboration d'une grammaire convoque la notion de « langue générale », et du même coup celle de corpus de référence, qui lui est souvent associée[28]. Un tel corpus relève à vrai dire de la gageure, mais il n'en demeure pas moins un idéal vers lequel il faut tendre. Biber, entre autres, a souligné le caractère artificiel d'une langue générale conçue comme une entité. En effet, la variation sous-tend la langue, qui correspond à un ensemble hétérogène de différents registres dont il convient de dégager les caractéristiques. En envisageant un corpus susceptible de représenter la langue générale, M.-P. Péry-Woodley fait remarquer à juste titre que la variation langagière est un obstacle difficile à surmonter : « le corpus équilibré est sans doute celui qui a 'de tout un peu', mais encore faudrait-il savoir ce qu'est 'tout', c'est-à-dire quelles sont les classes à représenter -ce qui nécessite un modèle complet de la variation-, et avoir accès à des textes les représentant » (1995 : 218). Pour des raisons analogues, B. Habert (2000) propose de restreindre le concept de langage à celui d' « emplois déterminés d'une langue », en raison de « notre ignorance de la population d'événements que constitue un langage dans son ensemble »[29]. Une telle réserve est d'autant plus justifiée pour la langue ancienne que nous n'aurons jamais accès à certains registres, en particulier ceux de l'oral, et que nous ignorons peut-être l'existence de certains : toute maîtrise de la variation est d'emblée exclue, et du même coup, plus encore que pour toute langue moderne, la perspective d'un corpus de référence au sens strict du terme. Cela n'empêche cependant pas de viser un corpus aussi représentatif que possible des états de langue successifs[30] que nous savons avoir existé (en gardant à l'esprit

[28] Pour une discussion de cette notion, généralement mise en relation avec celle de « corpus de spécialité », voir, entre autres, les différents travaux de D. Biber, de B. Habert, ainsi que Péry-Woodley (1995) et Condamines (2000).

[29] La question est en outre discutée dans Habert et Zweigenbaum (2002).

[30] « Etat de langue » s'entend ici prioritairement dans un sens chronologique : un état de langue à une époque donnée. Mais un état de langue est complexe, soumis aux variations liées aux genres et aux registres… qui constituent aussi d'une certaine manière, et dans un sens quelque peu différent, des « états de langue ».

qu'un corpus ne représentera jamais avec une entière fiabilité que la langue des textes qui le constituent).

Mais comment constituer un tel corpus dans la perspective d'une grammaire ? Un exemple nous est fourni pour l'anglais moderne par la *Longman Grammar of Spoken and Written English* (Biber *et al.*, 1999/2004), première grammaire qui affiche clairement le désir de dépasser les seules considérations structurelles pour envisager comment elles sont réellement utilisées dans le discours oral et écrit. Les auteurs se sont pour cela appuyés sur un corpus de 40 millions de mots, qui correspondent à 37244 textes, appartenant à 4 registres majeurs (divisés en sous-registres) : conversation, fiction, presse et prose académique. Ces registres, définis en termes non linguistiques (domaine, but, sujet…) sont considérés comme à la fois productifs et assez différents, et donc supposés couvrir la plupart de la variation en anglais[31]. Nous n'entrerons pas dans le détail de la constitution et de l'utilisation du corpus dans cette grammaire[32] : nous en retiendrons ici deux aspects majeurs. Il s'agit d'une part de l'établissement de données quantifiées, et d'autre part de l'exploitation à géométrie variable des corpus, en fonction des points à étudier : si certains dénombrements ont porté sur l'ensemble du corpus, d'autres analyses n'ont été faites que sur des extraits ou sur certains registres. Il est précisé en début de chaque chapitre quel corpus a été utilisé.

Le projet de *Grande grammaire historique du français*[33], qui a démarré récemment, diffère à bien des égards de la *Longman Grammar* : c'est un travail sur des états de langue ancienne, couvrant une diachronie de douze siècles ! Ces deux spécificités ajoutent des contraintes et des difficultés pour

[31] Ont cependant été ajoutés 2 registres : discours non conversationnel (discours, sermons…) et prose non fictionnelle, et 2 sous-corpus : news et conversation en anglais américain.

[32] Pour une présentation détaillée des objectifs et de la constitution du corpus, voir l'introduction, pp. 1-45. Sont aussi explicités les différents types d'analyses menées sur le corpus, automatiques pour certaines, partiellement manuelles dans le cadre de programmes interactifs pour d'autres.

[33] Projet sous la direction de C. Marchello-Nizia, B. Combettes, S. Prévost et T. Scheer.

la constitution d'un corpus représentatif. Parvenir à concevoir - et à réaliser- un corpus aussi satisfaisant que possible n'est pas simple. Il faut dans un premier temps déterminer ce qu'il faudrait « idéalement », non seulement du point de vue de la chronologie, et cela pose la question cruciale de la périodisation, mais aussi du point de vue des genres (ou des registres) et des dialectes. Il s'agit ensuite de rassembler les textes pertinents, or la tâche n'est pas simple en raison des problèmes d'existence ou de disponibilité des textes évoqués plus haut. Il reste enfin à faire en sorte que le corpus constitué, tout en étant représentatif, reste néanmoins « gérable », non seulement du point de vue de l'exploitation des textes (certains faits sont plus faciles à repérer automatiquement que d'autres), mais aussi du traitement des données extraites. Cela ne va pas de soi au regard de la diachronie retenue, et l'utilisation d'un corpus à géométrie variable est inévitable.

La question de la périodisation est particulièrement complexe. Faut-il envisager une périodisation relative aux faits étudiés, et dans ce cas comment articule-t-on la sélection des textes ? Ou bien faut-il imaginer une périodisation globale (avec d'éventuelles « exceptions) ? Il est possible de trouver un compromis en dépassant le traditionnel découpage en ancien français, moyen français… : en s'appuyant sur quelques critères minimaux, on peut en effet proposer une chronologie affinée de ces périodes, qui prenne aussi en compte des spécificités « dialectales », telles que l'anglo-normand au 12ème, ou le picard au 13ème. Il convient cependant d'œuvrer avec beaucoup de prudence dans la mesure où tout « découpage » donne un sens à l'évolution : c'est toujours une interprétation et/ou un présupposé théorique.

En ce qui concerne la diversification qualitative du corpus, se pose la question de s'en tenir aux domaines (religieux, littéraire, juridique, historique et didactique) ou d'affiner la classification et de pendre en compte les genres[34]. A cet égard il faut garder à l'esprit, d'une part la difficulté qu'il peut y avoir à trouver pour telle période tel texte appartenant à

[34] Voir note 9 la classification proposée dans le cadre du projet mené à l'ENS-LSH de Lyon.

tel genre, et d'autre part le fait que tous les genres n'existent pas à toutes les époques.

Au regard de ces différentes contraintes et spécificités, certains choix ont été faits. Précisons tout d'abord qu'il ne s'agit en aucun cas de tout réécrire : de nombreuses études excellentes existent, et il est inconcevable de ne pas les utiliser. Il faut cependant rester vigilant quant à l'exploitation des données quantifiées, s'il y en a, en précisant les sources et, autant que possible, les modalités d'analyse. Pour les études (partiellement) inédites, il est prévu l'exploitation d'un corpus à géométrie variable en fonction des phénomènes à étudier. On s'appuiera d'une part sur un gros corpus, soit pour puiser des exemples, soit pour effectuer des relevés simples et sans véritable analyse ensuite, et on utilisera d'autre part plusieurs sous-corpus (dont il reste à préciser le nombre et les modalités de constitution) pour des analyses quantifiées sur des aspects plus complexes. A l'instar de ce qui est fait dans la *Longman Grammar*, le corpus sur lequel s'appuie telle description ou telle analyse sera toujours explicité.

Il convient de veiller, pour la constitution de l'ensemble de ces corpus, au meilleur équilibre entre représentativité maximale au regard du « but » du corpus, et maniabilité. Le corpus le plus satisfaisant reste une articulation entre l'idéal, le disponible et le gérable.

2. Les apports

Le travail « sur corpus », c'est-à-dire sur des textes numérisés, suffisamment nombreux et diversifiés, permet la massification des données : cela est dû à la relative simplicité de leur extraction (certes variable selon le caractère plus ou moins complexe de l'expression recherchée, de son caractère plus ou moins abstrait) ainsi qu'à la facilitation de leur traitement (quant à elle variable en fonction du type d'analyse requis et des outils dont on dispose). Cette massification des données de travail permet un enrichissement conséquent de nos connaissances, tant du point de vue quantitatif que qualitatif, la combinaison des deux nous conduisant à une meilleure

compréhension des faits langagiers. C'est particulièrement essentiel pour une langue ancienne.

2.1. La quantification des données

Le fait de travailler sur des corpus numérisés facilite la détermination des fréquences des faits linguistiques. Dans le meilleur des cas, un simple « clic » permet de recenser le nombre d'occurrences de telle ou telle expression dans des corpus de plusieurs millions de mots. Outre la fréquence totale des occurrences, il est facile aussi d'en connaître la répartition dans les différents textes qui constituent le corpus.

Ces simples opérations, à elles seules, permettent de faire émerger des informations intéressantes, parfois inattendues. L'étude de deux marqueurs de topicalisation, *quant à X* et *à propos de X*[35] nous a ainsi réservé quelques surprises. La recherche portant sur les périodes du moyen français et du 16ème siècle, nous avons utilisé les bases du DMF et de Frantext-16ème, dans leur intégralité, sans donc envisager de pondérer la répartition des textes : craignant que les occurrences soient peu fréquentes, il avait été décidé de ne pas restreindre le nombre de textes. De fait, la fréquence de l'une des formes, *à propos de X*, s'est révélée très basse, et l'écart avec l'autre expression a largement dépassé nos prévisions. Nous avons en fait effectué deux dénombrements, le premier portant sur les occurrences en toutes positions, et le second sur celles en position initiale, en outre restreintes à la fonction de marqueur de topicalisation[36]. Nous avions prévu que l'écart serait important pour la période du moyen français, l'expression *quant à X* étant bien plus ancienne et donc susceptible d'être plus fréquente. C'est effectivement le cas : toutes positions confondues, on a 52 occurrences de *à propos de X* contre 1121 de *quant à X*[37], et pour la seule position initiale, on a 29 occurrences (7,8%) pour la première, contre 340 pour la

[35] Prévost (2008b)

[36] Ce qui a supposé un élagage « manuel » assez important pour *quant à X*.

[37] Soit 4,4% de l'ensemble des occurrences des deux formes pour le premier, et 95,6% pour le second.

seconde. On aurait pu penser que l'écart se serait réduit au siècle suivant, or ce n'est nullement le cas… au contraire. Toutes positions confondues, il n'y a plus que 19 occurrences (1,5%) de *à propos de X*, contre 1250 de *quant à X*. L'écart se creuse pareillement en position initiale : 11 occurrences (1,8%) contre 690 !

La répartition dans les textes des occurrences en position initiale[38] constitue un autre aspect intéressant. Elle montre en effet une évolution assez nette en ce qui concerne *à propos de X* : concentrée en moyen français dans quelques textes et chez de rares auteurs, l'expression, pourtant moins fréquente en valeur absolue au siècle suivant, apparaît dans davantage de textes et chez plus d'auteurs. De tels éléments auraient évidemment pu être relevés « à la main », mais le support numérisé permet de collecter en peu de temps un nombre élevé d'occurrences et de déterminer leur fréquence, permettant ainsi de se consacrer plus rapidement à l'analyse elle-même.

La couverture d'un vaste corpus est donc particulièrement précieuse quand on travaille sur des faits rares. Nous venons de le voir avec *à propos de X*, expression pour laquelle il aura fallu, pour le moyen français, un corpus de 7 milllions de mots pour extraire 29 occurrences en position initiale ! Cela reste peu, mais on ne peut en faire grief à la taille trop réduite du corpus : seule la rareté de l'expression est en cause. Même si le nombre d'occurrences sur lequel nous avons travaillé est faible, il a néanmoins permis de dégager certaines caractéristiques, voire des régularités, ce qui n'aurait pas été possible avec seulement quelques rares occurrences. Or étant donné leur forte concentration (chez seulement 3 auteurs, avec en outre 26 des 29 occurrences chez Christine de Pizan), une réduction mal ciblée du corpus (par exemple l'exclusion de ces trois auteurs) aurait tout simplement fait conclure à la non existence, à cette époque, de l'expression. Ce qu'aucune intuition de locuteur ne serait venu corriger.

[38] Nous n'avons pas regardé de près ce qu'il en est pour les occurrences toutes positions confondues.

D'une manière générale, plus le corpus est important et diversifié, et donc représentatif, mieux on peut dater avec exactitude l'émergence ou la disparition d'une forme. On peut aussi, et c'est un point essentiel, mesurer l'évolution de sa productivité. C'est particulièrement important pour les changements qui relèvent de la grammaticalisation. En effet, on y observe très souvent des phénomènes de réanalyse syntaxique ou de réinterprétation sémantique, qui se produisent d'abord dans des contextes discursifs spécifiques, avant que, selon un principe d'extension analogique, la nouvelle valeur ou forme s'étende à de nouveaux contextes, qui n'étaient initialement pas possibles. Souvent même la nouvelle valeur ou forme se substitue à l'ancienne, mais jamais brutalement. Cela se produit généralement après une phase de coexistence, la forme/valeur ancienne, d'abord majoritaire, cédant progressivement la première place[39]. Il est particulièrement intéressant de pouvoir quantifier avec précision ces différents changements, et au-delà des phénomènes individuels, de comparer de ce point de vue différentes évolutions, afin de mettre au jour, pour les caractériser, les différents rythmes auxquels se déroulent ces processus.

2.2. L'analyse qualitative

La simple attestation et quantification de faits permet de mettre au jour des éléments importants pour notre connaissance de la langue, mais elle est aussi le point de départ à une analyse qualitative des phénomènes. La quantité des données brassées permet en effet de donner davantage de poids et de sens à la mise en relation de « faits », de quelque niveau que ce soit. On peut ainsi, pour l'étude d'un phénomène spécifique, corréler certains « traits ». Dans le cadre de l'étude des énoncés à sujet pronominal de 3ème personne, nous avions fait l'hypothèse que la position du sujet était en partie corrélée au type d'éléments initiaux. Même s'il convient d'élargir le corpus, le traitement déjà effectué de 1247 énoncés à sujet pronominal (dont 173 à

[39] Parmi les approches récentes à ce sujet, voir Marchello-Nizia (2006, chap. 8).

sujet postverbal) a permis de confirmer l'hypothèse de départ, tout en l'affinant. Ainsi, les corrélations ont bien lieu entre telle position du sujet et telle expression initiale, mais elles mettent aussi en jeu des types d'éléments, de nature et de fonction assez diverses d'ailleurs : objets directs, adverbiaux temporels... On constate aussi que à nouveau une répartition se fait, au sein d'une même catégorie, entre les différentes expressions. Par exemple, on ne trouve de subordonnée causale que dans les énoncés à sujet préverbal, mais dans les énoncés à sujet postverbal, on peut parfois rencontrer des SN prépositionnels à valeur causale (*por ce*). Il est intéressant de noter que ces « répartitions complémentaires » ne se réalisent pas de la même manière d'un texte à l'autre : on observe des micro-systèmes propres à chaque texte, avec néanmoins ce dénominateur commun que sont les répartitions complémentaires[40].

Contrairement aux recherches pour lesquelles la collecte rapide de milliers d'occurrences se fait en quelques secondes (relevés des occurrences de *quant à X* et des différentes formes de *aucun*, par exemple), l'apport d'un corpus numérisé a été moins spectaculaire dans l'étude des énoncés à sujet pronominal. La recherche restait d'ailleurs envisageable en travaillant sur des textes non numérisés, et sans aucun outil de traitement informatique en aval, mais l'existence du support informatique a largement facilité et accéléré les tâches, permettant ainsi de se concentrer sur l'analyse et de réaliser l'étude en un temps raisonnable. La numérisation des textes a en effet permis une extraction rapide des énoncés à sujet pronominal (même s'il a fallu ensuite intervenir manuellement pour éliminer les occurrences non désirées, cependant réduites grâce à des requêtes d'extraction ciblées au mieux ; reste qu'un corpus annoté quant au type de propositions aurait été utile). L'extraction a été facilitée aussi pour certains des éléments initiaux de la phrase. L'étiquetage morpho-syntaxique dont bénéficiait l'un des textes retenus, précieux pour de nombreuses autres études, a moins apporté pour cette étude spécifique : *il* et

[40] Il serait intéressant de savoir si cela est spécifique au palier textuel, ou bien si des distributions analogues s'observent à des niveaux supérieurs, en particulier ceux de l'auteur et du genre.

ses variantes graphiques sont facilement repérables, sans qu'il soit nécessaire de passer par l'intermédiaire de l'étiquette « pronom personnel » ; pour les éléments initiaux c'est un étiquetage d'une autre type (positionnel, fonctionnel, voire sémantique…) qui aurait au contraire été utile. Nous reviendrons plus loin sur la question de l'enrichissement des textes.

Si la massification des données permet de rendre significatives des mises en relation de traits, qui en nombre trop réduit ne le seraient pas, elle permet aussi d'établir des corrélations entre différents phénomènes linguistiques isolés et de mettre ainsi au jour des macro-évolutions. Nous illustrerons ce fait par un exemple qui s'appuie sur les travaux de C. Marchello-Nizia.

Les démonstratifs se répartissaient en ancien français en deux séries, -« cist » et « cil »-, qui s'opposent sur le plan sémantique, la nature de cette opposition évoluant en moyen français pour devenir morphologique. Quelle que soit l'interprétation exacte que l'on fait de l'opposition sémantique d'origine (présence/absence d'une référence explicite au contexte immédiat ou à la situation d'énonciation, appartenance à/exclusion de la sphère du locuteur), on peut interpréter cette évolution comme un mouvement du subjectif vers l'objectif[41]. Un tel mouvement est intéressant en soi, mais il l'est d'autant plus qu'il peut être mis en relation avec celui de deux autres formes. Il s'agit d'une part du verbe *cuidier* (« croire »), qui, assez fréquent en ancien français, va disparaître entre le 15ème et le 17ème siècle. Or ce verbe dénote une implication forte du locuteur-énonciateur, qui se pose comme le détenteur de la vérité ou de la fausseté de ce qu'il asserte. Il s'agit d'autre part de l'apparition de l'adjectif *reel/real* (tout d'abord dans des textes juridiques), qui va désormais prendre en charge la « réalité », c'est-à-dire la vérité assertée comme objective, alors que celle-ci était jusque là dénotée par l'adjectif *verai/voir*, celui-là même qui exprime l'idée de vérité assumée comme

[41] Cf. Marchello-Nizia (1997b) pour la corrélation des trois phénomènes qui vont être évoqués, et Marchello-Nizia (2003), (2004a), (2004b) et (2006) pour la poursuite de l'analyse des démonstratifs.

subjective : la séparation entre objectif et subjectif est désormais entérinée. C. Marchello-Nizia (1997b) propose de voir dans ces trois changements une présence amoindrie du locuteur-énonciateur-sujet, et, donc une objectivation de la langue, évolution de nature méta-sémantique.

Dans le domaine de la grammaticalisation, la mise en relation de phénomènes convergents s'est révélée particulièrement fructueuse, permettant de mettre au jour des macro-évolutions, que la masse des données prise en compte rend incontestables.

Il ne faut cependant pas non plus surestimer le rôle de l'exploitation des corpus dans la mise au jour de ces phénomènes : beaucoup étaient connus, et malgré la difficulté pour un observateur à repérer les faits de corrélation, l'accumulation des études sur la langue ancienne avait permis d'en remarquer certaines. Elles n'étaient pas nécessairement formulées en termes de « macro-évolution », mais de ce point de vue c'est autant, ou même plus, une approche théorique différente qui a modifié le regard. Il reste que l'accès à des données massives et diversifiées, ainsi que la facilitation de leur traitement, a permis de quantifier et comparer les évolutions convergentes, et donc d'affiner leur description. Il a par ailleurs été possible de donner plus de poids aux faits assertés en établissant des chronologies assez précises.

2.3. Les corpus enrichis

Le fait de bénéficier d'un corpus enrichi constitue un apport considérable, encore trop rare pour la langue ancienne, même si la situation est en train de changer (cf. note 15).

Cela permet en premier lieu de simplifier et d'accélérer considérablement l'extraction des données dès qu'on travaille, non pas sur des formes brutes, mais sur des catégories, quelles qu'elles soient (pourvu qu'elles soient identifiées dans le texte)[42]. Or l'analyse de catégories, et non plus de simples

[42] Il faut à cet égard garder à l'esprit que travailler sur des corpus enrichis suppose une étape préalable de « familiarisation » avec l'étiquetage mis en œuvre, tant en ce qui concerne ses principes que sa réalisation dans les faits : un étiquetage n'est jamais neutre ni parfait, et il convient donc

formes brutes, permet d'accéder à une autre couche de description. La démarche rejoint de ce point de vue la mise au jour de macro-évolutions, en ce que dans les deux cas on ne raisonne plus simplement sur des formes en surface, mais on s'attache à des catégories morphologiques, des fonctions syntaxiques, ou des étiquettes sémantiques. La mise au jour de macro-évolutions et le travail sur des formes enrichies sont d'ailleurs souvent complémentaires : le fait de raisonner sur des catégories permet en effet non seulement de déceler des faits inédits, mais aussi de les mettre en relation. Nous mentionnerons à ce propos un exemple que nous avons déjà évoqué ailleurs (Prévost et Heiden 2002, Prévost 2005), mais que son exemplarité nous incite à mentionner à nouveau. La BFM renferme cinq textes étiquetés morphologiquement[43], deux d'ancien français (13ème siècle[44]), et trois de moyen français (début du 15ème pour l'un, fin du 15ème pour les deux autres[45]). La seule observation de la fréquence des différentes catégories révèle que, en ancien français, la catégorie la plus représentée est le verbe conjugué, alors que c'est le nom commun à partir du 15ème siècle, y compris dans les deux textes d'un genre identique (roman/nouvelle) à ceux de la période précédente. Bien que la prudence soit requise dès qu'il s'agit de généraliser (d'autant qu'il ne s'agit que de cinq textes, et tous littéraires), il est bien tentant de s'y aventurer dans la mesure où le phénomène a été pareillement constaté dans le cadre d'un étiquetage par apprentissage. Appliqué sur un texte d'ancien français (*La Queste del Saint Graal*), l'étiqueteur a proposé comme première règle d'étiquetage des formes inconnues : « tout mot contenant le caractère « e » est à étiqueter 'verbe conjugué' ». Le même étiqueteur, entraîné sur un corpus de

d'en maîtriser les caractéristiques et les faiblesses (zones d'erreurs fréquentes en particulier), afin d'exploiter de manière appropriée le corpus.

[43] Pour une présentation du jeu de 58 étiquettes, voir <http ://weblex.ens-lsh.fr/projects/bfm/jeu_d_etiquettes>.

[44] Il s'agit des romans *La Mort Artu* et *La Queste del Saint Graal.*

[45] Respectivement *Les quinze joyes de mariage*, recueil de nouvelles, et *Le roman de Jehan de Paris* et le livre 1 des *Mémoires* de Commynes.

français moderne (mêlant des genres différents), a fourni, comme même première règle : « tout mot contenant le caractère « e » est à étiqueter 'nom commun' ».

L'un des aspects frappants dans cette découverte est le décalage entre la relative simplicité de la démarche -dénombrement de catégories- et le résultat obtenu (dont il faudrait évidemment approfondir l'analyse). Si une démarche aussi simple permet d'obtenir des informations aussi intéressantes, on peut légitimement imaginer qu'une exploitation plus fine permettra des découvertes d'une grande importance.

On peut à cet égard envisager un type d'analyse qui, très développé désormais pour les langues modernes[46], l'est encore peu -ou pas- pour les langues anciennes. Il s'agit de l'analyse multidimensionnelle qui permet de mettre en évidence, à l'aide d'outils statistiques, des corrélations entre faits langagiers, qu'il s'agisse de formes brutes ou enrichies. Au-delà, comme nous l'avions évoqué ci-dessus, on peut envisager d'élaborer une typologie « inductive » des textes (c'est-à-dire fondée sur des traits linguistiques), avec la perspective d'établir une corrélation ultérieure entre critères « externes » (situationnels) et « internes » (linguistiques), ce qui permettrait la mise au jour d'une typologie affinée des textes, non encore réalisée pour un corpus de français médiéval[47].

Conclusion

Si le recours aux gros corpus numérisés constitue un apport précieux pour tout état de langue, c'est bien plus vrai encore lorsqu'il s'agit d'une langue (ou d'une succession d'états de langue) pour laquelle il n'existe plus de locuteurs. Massification raisonnée des textes, traitement automatisé des données, enrichissement croissant des textes : tout cela permet de découvrir, de corriger et d'enrichir notre connaissance des faits linguistiques, dont la mise en relation est en outre facilitée,

[46] Pour le français, voir Fleury & al. (2000), et Folch & al. (2000).

[47] Mais réalisée sur des états de langue moderne : voir Habert (2000) ainsi que les différents travaux de Biber (en particulier Biber 1998).

ce qui favorise la mise au jour de vastes mouvements d'évolution.

Afin de tirer pleinement parti de cet outil remarquable sans en subir les éventuels inconvénients, il est indispensable d'éviter les généralisations abusives, y compris lorsque les textes présentent des caractéristiques situationnelles similaires. En effet, en langue ancienne (au moins pour le français), on constate que les tendances stylistiques des auteurs peuvent parfois très fortement « typer » les textes. Cela prouve que le recours aux gros corpus et l'automatisation des traitements ne doit pas exclure une attention constante à l'éventuelle spécificité des résultats de chaque texte, dans le détail desquels il est par ailleurs nécessaire de « descendre » autant que possible. On risque sinon d'écraser certaines oppositions, de neutraliser certaines variations. Or ce sont elles précisément qui font toute la richesse de la langue : jamais la quantité des données ne doit évincer la qualité de leur analyse. C'est assurément un défi de taille, mais qui peut nous permettre d'accéder à la compréhension des états successifs de la langue ancienne, qui nous sont donnés à voir dans l'ensemble des textes qui constituent nos corpus.

Bibliographie

Biber, D. (1988). *Variation across speech and writing*. Cambridge : Cambridge University Press.

Biber, D. (1990). « Methodological issues regarding corpus-based analyses of linguistic variation ». *Literary and Linguistic Computing*, 5, (4), pp. 257-270.

Biber, D. (1993). « Using register-diversified corpora for general language studies ». *Computationnal linguistics*, 19, (2), pp. 219-241.

Biber, D. (1995). *Dimensions of register variation : a cross-linguistic comparison*. Cambridge : Cambridge University Press.

Biber, D., Conrad, S. & Reppen, R. (1998). *Corpus Linguistics. Investigating Language Structure and Use.* Cambridge : Cambridge University Press.

Biber, D., Johansson, S., Leech G., Conrad S., Finegan E., (2004, 1ère éd. 1999) *The Longman Grammar of Spoken and Written English*, London : Longman.

Guillot, C. (2003). *Le rôle du démonstratif dans la cohésion textuelle au 15ème siècle. Eléments de grammaire textuelle*. Thèse de doctorat. Lyon : Ecole normale supérieure Lettres et Sciences humaines.

Fleury S., Folch H., Habert B., Heiden S., Illouz G., Lafon P. et Prévost S. (2000).« Profilage de textes : cadre de travail et expérience ». In *Actes du colloque 'JADT 2000 : 5es Journées Internationales d'Analyse Statistique des Données Textuelles', Lausanne, 2000*, vol 1, pp. 163-170

Folch H., Heiden S., Habert B., Fleury S., Illouz G., Lafon P., Nioche J. et Prévost S. (2000). « TyPTex : Inductive typological text classification by multivariate statistical analysis for NLP systems tuning/evaluation ». In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, G. Stainhaouer (éds), *Second International Conference on Language Resources and Evaluation*, pp. 141-148

Habert, B., Nazarenko, A. & Salem, A. (1997). *Les linguistiques de corpus*. Paris : A. Colin / Masson.

Habert, B. (2000). « Des corpus représentatifs : de quoi, pour quoi, comment ? ». In M. Bilger (éd), *Cahiers de l'université de Perpignan, 31, 'Linguistiques sur corpus. Etudes et réflexions'*, Perpignan : Presses universitaires de Perpignan, pp. 11-58.

Habert, B & Zweigenbaum, P. (2002). « Régler les règles ». *TAL*, 43, (3), pp. 83-105.

Heiden, S. & Lavrentiev, A. (2004). « Ressources électroniques pour l'étude des textes médiévaux : approches et outils ». *Revue Française de Linguistique Appliquée*, IX, (1), pp. 99-118.

Heiden, S. & Guillot, C. (2003). « Capitalisation des savoirs par le Web : une application de la TEI pour l'encodage et l'exploitation des textes de la Base de Français Médiéval ». In P. Kunstman, F. Martineau & D. Forget (éds), *Ancien et moyen français sur le web. Enjeux méthodologiques et analyse du discours*, Ottawa : Les Editions David, pp. 77-92.

Kunstmann, P. (2000). « Ancien et moyen français sur le web : textes et bases de données ». *Revue de Linguistique Romane*, 64, pp. 17-42.

Lightfoot, D. W. (1979). *Principles of diachronic Syntax*. Cambridge : Cambridge University Press.

Marchello-Nizia, C. (1995). *L'évolution du français : ordre des mots, démonstratifs, accent tonique*. Paris : Armand Colin.

Marchello-Nizia, C. (1997a). *La langue française aux 14ème et 15ème siècles*. Paris : Nathan. (Première publication 1979 Bordas).

Marchello-Nizia, C. (1997b). « Evolution de la langue et représentations sémantiques : du 'subjectif' à l' 'objectif' en français ». In C. Fuchs & S. Robert (éds), *Diversité des langues et représentations cognitives*, Paris : Ophrys, pp. 119-135.

Marchello-Nizia, C. (1999). « Corpus diachroniques ». *Revue française de linguistique appliquée*, IV, (1), pp. 31-39.

Marchello-Nizia, C. (2000). « Les grammaticalisations ont-elles une cause ? », *L'information grammaticale*, 67, pp. 3-9.

Marchello-Nizia, C. (2003). « 'Se voz de ceste ne voz poéz oster, Je voz ferai celle teste coper' (*Ami et Amile* 753). La sphère du locuteur et la deixis en ancien français ». In A. Vanneste, P. de Wilde, S. Kindt et J. Vlemings (éds), *Memoire en temps advenir. Hommage à Theo Venckeleer*, Louvain : Peeters, p. 413-427.

Marchello-Nizia, C. (2004a). « La sémantique des démonstratifs en français : une neutralisation en progrès ? », *Langue française*,141, pp. 69-84.

Marchello-Nizia, C. (2004b). « Deixis and subjectivity: the semantics of demonstratives in old French (9th-12th century) », *Journal of Pragmatics*, 37 (1), p. 43-68.

Marchello-Nizia, C. (2006). « From personal deixis to spatial deixis: the semantic evolution of demonstratives from Latin to French ». In M. Hickman et S. Robert (éds), *Space in Languages: Linguistic Systems and Cognitive Categories*, Amsterdam/ Philadelphie : John Benjamins, pp. 103-120.

Pery-Woodley, M.-P. (1995). « Quels corpus pour quels traitements automatiques ? », *TAL*, 36, (1-2), pp. 213-232.

Prévost, S. (1999). « *Aussi* en position initiale : évolution sémantico-syntaxique du 12ème au 16ème siècle ». *Verbum*, XXI, (3), pp. 351-380.

Prévost, S. (2003). « *Quant a* : analyse pragmatique de l'évolution diachronique (14ème-16ème siècles) ». In B. Combettes, A. Theissen & C. Schnedecker (éds.), *Ordre et distinction dans la langue et le discours*, Paris : H. Champion, pp. 443-459.

Prévost, S. (2005). « Exploitation d'un corpus de français médiéval : enjeux, spécificités et apports ». In A. Condamines (éd), *Sémantique et corpus*, Paris : Hermès/Lavoisier, pp. 147-176.

Prévost, S. (2007). « *À propos de, à ce propos, à propos* : évolution du 14ème au 16ème siècle », *Langue Française*, 156, pp. 108-126.

Prévost, S. (2008a). « Evolution de la position du sujet pronominal dans quelques textes du 14ème siècle : un cas de grammaticalisation ? », *Journée d'études 'Language over time : A symposium on Anglo-Norman in the context of medieval French language use'*, Birmingham, janvier 2008.

Prévost, S. (2008b). « Evolution de quelques marqueurs de topicalisation du 14ème au 16ème siècle », *L'information grammaticale*, 118, pp. 38-43.

Prévost, S. & Heiden, S. (2002). « Etiquetage d'un corpus de français médiéval : enjeux et modalités ». In C.D. Pusch & W. Raible (éds.), *Romance Corpus Linguistics - Corpora and Spoken Language*, Tübingen : Gunter Narr Verlag Tübingen, pp.127-136.

Prévost, S. &. Schnedecker, C. (2004). « *Aucun(e)(s) /d'aucun(e)(s) / les aucun(e)(s)* : évolution du français médiéval au français moderne ». *Scolia*,18, pp. 39-73.

Sinclair, J. (1996). *Preliminary recommendations on Coprus Typology*. Technical report, EAGLES, (Expert Advisory Group on Language Engineering Standards). <http://www.ilc.cnr.it/EAGLES96/pub/eagles/corpora/corpustyp.ps.gz>

Stein, A. (2003). « Etiquetage morphologique et lemmatisation de textes d'ancien français ». In P. Kunstman, F. Martineau & D. Forget (éds), *Ancien et moyen français sur le web.*

Enjeux méthodologiques et analyse du discours. Ottawa : Les Editions David, pp. 273-284.